O FLUXO TEMPORAL DE TERMOS RELEVANTES: uma análise em teses da UFMG de 2007 a 2018 nas ciências humanas.

THE TEMPORAL FLOW OF RELEVANT TERMS: an analysis in UFMG theses from 2007 to 2018 in human sciences

Luiz Antônio Lopes Mesquita
Universidade Federal de Minas Gerais

Célia da Consolação Dias
Universidade Federal de Minas Gerais

Renato Rocha Souza
Fundação Getúlio Vargas
Universidade Federal de Minas Gerais

RESUMO

O objetivo geral desta pesquisa foi analisar se há uma variação temporal característica da distribuição de valores de termos relevantes ao longo do tempo da produção de textos que possa contribuir como um critério para o processo de sua indexação automática. Foram analisadas as teses de doutorado dos programas de pós-graduação (PPGs) da área de Ciências Humanas da UFMG, considerando-se 7 PPGs distintos, sendo cada um deles um corpus, com um total de 929 teses defendidas período de 12 anos, de 2007 a 2018. Os termos considerados foram todos os sintagmas nominais contidos nos próprios textos das teses. Cada sintagma nominal recebeu um valor associado à sua relevância como descritor de acordo com os critérios de frequência do termo na própria tese (*TF - Term Frequency*) e com o inverso da frequência de ocorrência do termo no total de teses de cada PPG (*IDF - Inverse Document Frequency*). As teses foram divididas em 12 grupos em cada PPG para o cálculo da data média de defesa das teses e da média de pontuação consolidada dos termos relevantes nas teses. Como resultados, identificou-se o comportamento característico de cada PPG através de um gráfico de dispersão do nível médio de pontuação de relevância ao longo do tempo. Para cada gráfico de cada um dos 7 PPGs foi adicionada uma linha de tendência, considerando seu respectivo R^2, e feita sua análise específica. Todos os comportamentos de distribuição temporais foram caracterizados em equações e podem ser aplicados como critério para indexação automática.

Palavras-Chave: Recuperação da Informação Temporal. Indexação Automática. Sintagmas Nominais.

ABSTRACT

This research's general objective was to analyze if there is a temporal variation characteristic of the distribution of values of relevant terms over the time of the production of texts that can contribute as a criterion for the automatic indexing process. The doctoral theses of the graduate programs (PPGs) in Human Sciences at UFMG were analyzed, considering seven different PPGs, each of which is a corpus, with 929 theses defended in a period of twelve years, from 2007 to 2018. The terms considered were all the noun phrases contained in the texts of the theses. Each noun phrase received a value associated with its relevance as a descriptor according to the term frequency criteria in the thesis itself (TF - Term Frequency) and with the inverse of the frequency of occurrence of the term in the total of theses of each PPG (IDF - Inverse Document Frequency). The theses were divided into 12 groups in each PPG to calculate the average defense date of the theses and the average consolidated score of the relevant terms in the theses. As a result, each PPG's characteristic behavior was identified through a scatter plot of the average level of relevance score over time. For each graph of each of the 7 PPGs, a trend line was added, considering its respective R^2, and its specific analysis was made. All temporal distribution behaviors were characterized in polynomial equations and applied as a criterion for automatic indexing.

Keywords: Temporal Information Retrieval. Automatic Indexing. Noun Phrase.

1. INTRODUÇÃO

Bush (1945, p. 95) já predissera o crescimento da informação e suas consequências ao pontuar sobre a necessidade da interdisciplinaridade:

> "O investigador fica pasmo com as descobertas e conclusões de milhares de outros pesquisadores - conclusões que ele não consegue encontrar tempo para apreender, muito menos para lembrar, tal como aparecem. No entanto, a especialização se torna cada vez mais necessária para o progresso, e o esforço para estabelecer uma ponte entre as disciplinas é ainda superficial" (ibidem, tradução livre).

Para Saracevic (1996, p. 42), Bush (1945), como cientista do MIT e em plena Segunda Guerra Mundial, não só aponta o problema da "explosão informacional" como também sua possível solução com o uso das "tecnologias da informação", criando o cenário para o surgimento da Ciência da Informação (CI) nos anos 50. Mooers (1951) aponta um dos caminhos da C.I. que ele denominaria como Recuperação da Informação (RI) através de seu protótipo Zatocoding.

A Ciência da Computação também "desenvolve significativas pesquisas nessa área [RI] com o objetivo principal de prover aos usuários de seus sistemas um fácil acesso à informação do seu interesse" (BAEZA-YATES; RIBEIRO-NETO, 2011, p. 1, tradução livre). Dentre muitas outras áreas que tornam a RI interdisciplinar, a Linguística contribui significativamente para o processamento de informações textuais em linguagem natural.

A contabilização de palavras isoladas pode ser feita facilmente com a identificação de delimitadores, como o espaço. Com os algoritmos de Processamento de Linguagem Natural (PLN) é possível utilizar cada vez mais estruturas linguísticas complexas. Uma dessas estruturas é o sintagma nominal (SN). Perini *et al.* (1996) apresentam que o SN possui maior valor semântico que a palavra isolada.

Os SNs podem ser extraídos automaticamente de textos. Os trabalhos de Kuramoto (1996), Souza (2005), Maia (2008), Corrêa *et al.* (2011), Mesquita et al. (2013, 2014) e outros apresentam como tema central a utilização de SNs através da sua extração em PLN de forma semiautomática e automática para a língua portuguesa. O crescente volume informacional somado à crescente capacidade de processamento, juntamente com avanços no PLN, abre espaço para novas pesquisas. Uma delas estaria no uso de sintagmas nominais em sistemas de recuperação da informação, como a indexação automática.

Os critérios para escolha de descritores para a indexação automática possuem suas raízes históricas nos conceitos de "frequência do termo", de Luhn (1957), e de "especificidade", de Sparck Jones (1972). Posteriormente a eles, vários outros critérios foram apresentados e podem ser atribuídos a 8 classes distintas (BORGES; LIMA, 2015). Para Mathews e Kanmani (2012), um dos critérios mais recentes estaria ligado ao que denominam como "Temporal

Information Retrieval" em virtude da quantidade imensa de dados disponíveis na Internet e que são fortemente dependentes do tempo.

Com base nas técnicas de "Recuperação da Informação Temporal", Duchon *et al.* (2015) apresentam, sob forma de patente, um método de conversão de dados e fluxos de tópicos, combinados com métodos temporais, para prever objetivamente atividades de tópicos no futuro.

O objetivo geral desta pesquisa foi analisar se há uma variação temporal característica da distribuição de valores de termos relevantes ao longo do tempo da produção de textos que contribui como um critério para o processo de sua indexação automática.

Este artigo está organizado em 5 seções: Introdução, Fundamentação Teórica, Metodologia, Resultados e Considerações Finais.

2. FUNDAMENTAÇÃO TEÓRICA

Kuramoto (1999) apresentou em sua tese de doutorado uma das primeiras aplicações para computador utilizando o SN para recuperação de informação na língua portuguesa. Souza (2005), a partir desses estudos, propôs uma metodologia de escolha automática de SNs como descritores relevantes no processo de indexação automática. Maia (2008) desenvolveu uma ferramenta , a partir da metodologia de Souza (2005), que, dentre outras funcionalidades, extrai os SNs de forma automática. Mesquita et al. (2013, 2014), a partir da ferramenta de Maia (2008), verificaram comportamentos da pontuação de descritores relevantes em textos científicos.

Os SNs em um documento apresentam densidade informacional superior à palavras isoladas, mantendo maior proximidade do discurso contido nos documentos por eles descritos (KURAMOTO, 1996; SOUZA, 2005). "Palavras isoladas, como descritores, podem apresentar mais problemas de polissemia ou de plurisignificação" (LYONS, 1987, p. 140). Além de apresentarem menos influência dos problemas acima, "os sintagmas nominais trazem em seu bojo o contexto semântico dos discursos" (SOUZA, 2005, p. 136). Para Baeza-Yates e Ribeiro-Neto (2011) os substantivos, que compõem um SN, possuem maior valor semântico ao serem usados como termos de indexação. Portanto, o uso de SNs como termos de indexação pode apresentar melhores resultados que o uso de palavras isoladas.

Para Cintra (1983, p. 9) um descritor pode ser analisado formalmente como um "sintagma de símbolos notacionais (números, letras, pontuação, marcas) isto é, unidades resultantes da combinação de formas menores em unidades de nível superior. Ex.: leit — eira ->• leiteira; o,

vestido, verde, de, Lúcia ->* O vestido verde de Lúcia". Este último é um exemplo de sintagma nominal.

A indexação pode ser definida como:

> "[...] o processo de analisar o conteúdo informacional dos registros do conhecimento e sua expressão na linguagem do sistema de indexação. Ele implica: a) Selecionar os conceitos indexáveis de um documento; e b) Expressar esses conceitos na linguagem do sistema de indexação" (BORKO, 1978, p. 8).

Além da inviabilidade do tratamento de grandes quantidades de documentos, os problemas práticos da atividade de indexação manual encontram-se também na inconsistência praticada pelos indexadores (NAVES; DIAS, 2007), que podem ser interindexadores e intraindexadores (BORKO, 1977). A inconsistência interindexadores ocorre quando dois ou mais indexadores elegem ou atribuem descritores diferentes para um mesmo documento. A inconsistência intraindexadores ocorre quando um mesmo indexador atribui descritores diferentes para um mesmo documento em momentos diferentes.

A indexação automática se justifica então pela sua capacidade de atender ao crescente volume de documentos eletrônicos e de forma mais consistente que a manual. As pesquisas em indexação automática ganharam força após a Segunda Guerra Mundial, quando o espírito pragmático e o apoio em pesquisa tecnológica dos Estados Unidos geraram um grande avanço, permitindo várias implementações (ORTEGA, 2009).

Os conceitos básicos para os modelos de recuperação da informação surgem com Luhn (1957) assumindo a frequência do termo (*term frequency - TF*) como critério para atribuição de pesos em um documento.

> **Definição**: *Frequência do Termo*. O valor, ou peso, de um termo k_i que ocorre em um documento d_j é simplesmente proporcional à frequência do termo $f_{i,j}$. Isto é, quanto mais o termo k_i ocorre em um texto do documento d_j, mais alto é seu peso por frequência de termo $TF_{i,j}$ (LUHN, 1957, tradução livre).

O nível de exaustividade adotado é considerado como a principal decisão da política de indexação e vai determinar estatisticamente a quantidade de termos de indexação usada em média para cada documento. Uma indexação exaustiva elege/atribui termos de indexação para todos os assuntos de um documento, por outro lado, a indexação seletiva elege/atribui uma quantidade limitada de termos de modo a representar somente os assuntos principais de um documento (LANCASTER, 2004).

A exaustividade ótima considera que o número de termos de indexação deva ser otimizado de modo que a probabilidade de relevância do documento recuperado seja maximizada (BAEZA-YATES; RIBEIRO-NETO, 2011). Ou seja, para uma provável consulta, a quantidade

de termos de indexação deve possibilitar uma máxima recuperação de documentos considerados relevantes por um usuário.

A especificidade é a propriedade semântica do termo que depende do seu significado. Por exemplo, moradia é menos específico que casa ou apartamento. A especificidade pode ser ainda definida através da estatística em substituição da propriedade semântica do termo de indexação. Ou seja, o valor de especificidade de um termo pode ser calculado através do inverso da quantidade de documentos nos quais ele ocorre. Se um termo ocorre em todos os documentos, sua especificidade é baixa ou nula.

Baeza-Yates e Ribeiro-Neto (2011, p. 74) apresentam três recomendações de equações para o cálculo de pesos para termos em um documento:

1. $f_{i,j} \cdot \log N/n_i$
2. $1 + \log f_{i,j}$
3. $(1 + \log f_{i,j}) \cdot \log N/n_i$

Nas fórmulas acima:

- $f_{i,j} \rightarrow$ frequência do termo i no documento j (TF);
- $N/n_i \rightarrow$ número total de documentos dividido pelo número de documentos nos quais ocorre o termo i ao menos uma vez (especificidade ou IDF).

A terceira equação acima foi utilizada na seção metodologia desta pesquisa.

Moulahi *et al.* (2016) defendem que, no contexto exponencialmente crescente de produção mundial de dados com informações temporais relacionadas, um campo específico na RI tornou-se oportuno, uma vez que a dimensão temporal tem sido amplamente explorada como um critério de relevância para a RI. A aplicação da dimensão temporal pode ser percebida em três níveis: (1) da consulta, (2) do conteúdo do documento e (3) na recuperação do documento. A dimensão temporal, como um tipo de RI, faria interseção com outras áreas da RI como: (1) RI tradicional na Web, (2) RI em redes sociais, (3) RI na web a partir de dispositivos móveis e (4) RI em informações geotemporais (MOULAHI; TAMINE; YAHIA, 2016, p. 729).

Esta pesquisa visa contribuir para a recuperação da informação temporal, usando as linhas de tempo (*timelines*) na sua forma de apresentação de resultados. Essa contribuição, por sua vez, pretende ser usada também como um critério para a extração de descritores para a indexação automática de textos científicos através dos sintagmas nominais contidos nesses textos em língua portuguesa. As pesquisas sobre sintagmas nominais para a recuperação da informação em textos em Português ainda são raras, sendo que esse tipo de pesquisa

considerando também a RI temporal se mostrou inédita na literatura até o presente momento.

3. METODOLOGIA

Em virtude da necessidade de um corpus com textos que caracterizassem um aspecto temporal, buscou-se por teses de doutorado, como textos concebidos por um período de tempo relativamente mais longo (cerca de 4 anos) e acessíveis digitalmente. Foi escolhido o Repositório Institucional da UFMG (RI-UFMG), uma vez que o mesmo, por se tratar da mesma instituição na qual foi desenvolvida essa pesquisa, permitiria avaliar mais informações transversais nas análises de dados.

Além do recorte do tipo de documento, também, para essa pesquisa, foi definido um recorte temporal de um total de 12 anos e relacionados ao período contínuo de 2007 a 2018. O ano inicial de 2007 está relacionado a uma portaria da CAPES de 2006 que passa a regular que teses e dissertações sejam disponibilizadas através de repositório digital. O limite superior do ano de 2018 foi utilizado pois os dados foram coletados em 2019, sendo que as publicações deste mesmo ano ainda estavam sendo publicadas.

Ainda para justificar o recorte, foram considerados somente os PPGs que tiveram ao menos 12 teses no referido recorte temporal anterior. Essa limitação mínima quantitativa é referente a uma possibilidade de análise de dados considerando ao menos uma tese por cada um dos 12 intervalos temporais. Para esta publicação foi escolhida a área de Ciências Humanas. Foram analisadas 929 teses em 7 PPGs distintos. A Tabela 1, a seguir, apresenta o quantitativo total por PPG, sendo cada um deles considerado como um *corpus* para esta pesquisa.

Tabela 1 - Quantidade de teses analisadas por PPG (corpora).

Programa de Pós-Graduação	**Teses**	**%**
GRANDE ÁREA - CIÊNCIAS HUMANAS	929	100,0%
1 - Educação - Conhecimento e Inclusão Social	513	55,2%
2 - História	112	12,1%
3 - Geografia	100	10,8%
4 - Filosofia	80	8,6%
5 - Ciência Política	61	6,6%
6 - Sociologia	51	5,5%
7 - Ciências Humanas Sociologia e Política	12	1,3%

A metodologia consistiu essencialmente em três etapas, cujas partes são descritas a seguir:

1. Extração dos SNs:

1.1. Obtenção dos documentos originais (em formato PDF);

1.2. Conversão dos documentos originais para o formato texto (TXT);

1.3. Etiquetagem e extração dos SNs usando o Palavras ;

1.4. Tratamento dos SNs (retirada das partes determinantes[1], e a retirada de SNs de acordo com uma *stoplist*);

2. Seleção e pontuação dos SNs candidatos a descritores em cada tese:

2.1. Processamento do atributo de frequência dos SNs por documento (*TF*);

2.2. Processamento do atributo de especificidade dos SNs por *corpus* (*IDF*);

2.3. Pontuação da relevância ($P_{Tese\text{-}SN}$) dos SNs (através do *TF* e *IDF*) utilizando a terceira fórmula citada acima;

3. Distribuição da pontuação dos SNs candidatos a descritores:

3.1. Consolidação da pontuação dos SNs: nessa etapa, todo SN em cada tese recebeu uma pontuação $P_{Tese\text{-}SN}$ de acordo com seu TF/IDF. Em cada tese, a pontuação $P_{Tese\text{-}SN}$ de seus SNs foram somadas ($P_{Tese\text{-}Soma\text{-}SNs}$).

3.2. Normalização da pontuação dos SNs: para evitar que teses com maior número de páginas gerassem um viés nos dados, o valor de $P_{Tese\text{-}Soma\text{-}SNs}$ de cada tese foi reduzido a 1 dividindo-se $P_{Tese\text{-}Soma\text{-}SNs}$ por ele mesmo. Logo, cada pontuação de cada SN na tese, $P_{Tese\text{-}SN,}$ também foi dividido por $P_{Tese\text{-}Soma\text{-}SNs}$, resultando na pontuação do sintagma nominal normalizada, $P_{Tese\text{-}SN\text{-}Normalizada.}$

3.3. Pontuação do SN no PPG: após a normalização, cada SN recebeu uma pontuação geral para o PPG somando-se as respectivas pontuações normalizadas em cada tese, $P_{Tese\text{-}SN\text{-}Normalizada}$, e resultando na sua pontuação geral no PPG, $P_{PPG\text{-}SN}$.

3.4. Distribuição da pontuação geral de cada SN no PPG: a pontuação geral no PPG de cada SN, $P_{PPG\text{-}SN,}$ foi redistribuída de acordo com o TF de cada SN em cada tese, resultando na $P_{Tese\text{-}SN\text{-}Normalizada.}$

[1] Os determinantes são os artigos, os pronomes possessivos, os pronomes demonstrativos, os adjetivos interrogativos, relativos e indefinidos e, ainda, os numerais que constituem o SN e dependem do substantivo, cabeça ou constituinte principal do SN (DUBOIS *et al*, 2013, pg. 180).

3.5. Consolidação final por tese no PPG: após a redistribuição da pontuação normalizada de cada SN, $P_{Tese\text{-}SN\text{-}Normalizada}$, foi possível calcular a pontuação acumulada em cada tese somando-se esses mesmos valores em cada tese, $P_{PPG\text{-}Tese.}$

3.6. Consolidação por dimensão temporal: cada tese foi posicionada por ordem da data de sua defesa. As teses foram divididas igualmente (exceto por necessidade de arredondamento) em 12 partes. Cada uma das 12 fatias de tempo recebeu uma data calculada pela média das datas de defesas das suas respectivas teses, assim como recebeu uma média das pontuações obtidas em cada tese, $P_{PPG\text{-}Tese.}$

3.7. Caracterização das distribuições de pontuações por funções matemáticas: foram utilizadas técnicas de regressão, para caracterizar o comportamento da distribuição temporal dos valores de pontuações médias em cada PPG.

A seguir são apresentados os gráficos de comportamento da distribuição de pontuações por dispersão temporal, assim como suas equações de regressão polinomial e a análise desses resultados.

4. RESULTADOS

A seguir são apresentados os valores de pontuação para cada PPG, sendo estes distribuídos por dispersão com base nas datas médias em cada uma das suas 12 fatias temporais. Horizontalmente, a proximidade de pontos representa que houve uma maior densidade de defesas de teses em tal momento. Um maior espaçamento, significa o contrário.

Para sistematizar a análise dos gráficos, foram utilizadas regressões com seu respectivo R^2, que representa, numa variação de 0 a 1, o quanto a regressão passa próxima aos pontos originais (1 para o melhor caso). Usualmente considera-se acima de 0,95 uma boa regressão, embora esse valor possa esconder outras possíveis regressões, como a exponencial, por exemplo.

A seguir são apresentadas duas seções, sendo a primeira uma análise detalhada individual de cada PPG e a segunda análise geral das Ciências Humanas.

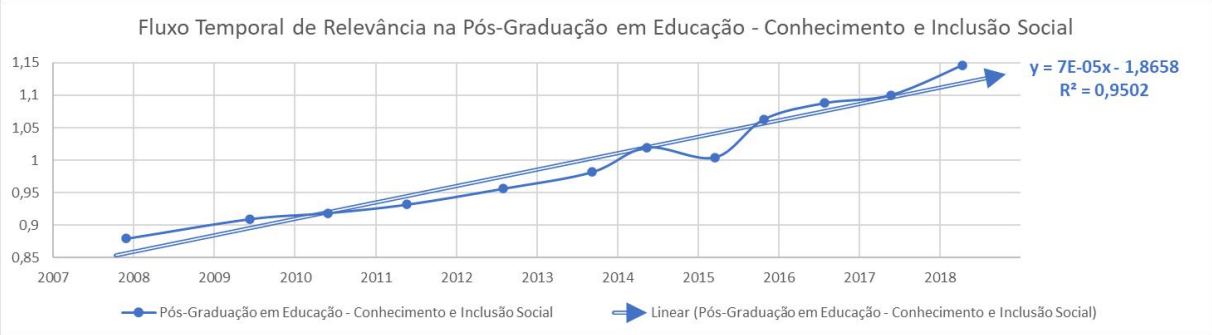
4.1. Análise de distribuição temporal por PPG por regressão polinomial

As curvas de cada PPG com suas respectivas linhas de tendência, obtidas por regressão, são apresentadas a seguir. Para cada PPG foi testada uma regressão linear simples e cinco polinomiais (variando de grau 2 ao 6). Para cada PPG foi testada à regressão mais simples e que obteve o R^2 satisfatório para as condições de análise aqui empregadas. Alguns trechos foram segmentados de modo a encontrar uma expressão que característica para um período. A seguir, juntamente com os gráficos, são apresentadas para cada PPG da grande área da Ciências Humanas as regressões escolhidas com base nesses critérios.

4.1.1. PPG – Educação - Conhecimento e Inclusão Social.

No Gráfico 1 abaixo, é possível dividir o fluxo temporal em três momentos (no início e antes de 2010, quando estão acima da linha de tendência, ao meio, quando estão abaixo desde 2010 até início de 2015, e a partir do final de 2015, quando voltam a ficar acima da linha).

Gráfico 1 - Fluxo Temporal de Relevância na Pós-Graduação em Educação - Conhecimento e Inclusão Social.



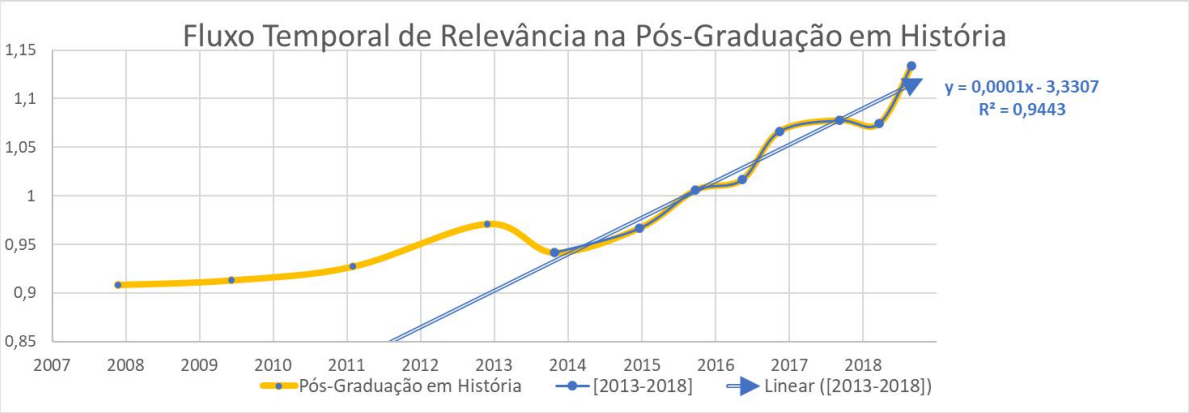
A regressão linear no Gráfico 1 acima também é expressa através de uma equação matemática à direita do final da linha. Nela existem dois valores importantes: o coeficiente "a", que é o número que antecede x, e o R^2, que indica o quão próximos ficaram os pontos da reta. O R^2 varia de 0 a 1, e seu valor máximo significa que a reta passou exatamente sobre todos os pontos, quanto mais eles se afastarem da linha da regressão, menor será o R^2.

Embora possa parecer sutil, o PPG em "Educação - Conhecimento e Inclusão Social" apresenta um significativo aumento no seu último valor correspondente a 2018, o que indica, em relação aos demais pontos, aquele que mais se destacou, tanto pelo seu valor absoluto de relevância, como também por se distanciar positivamente da referência da regressão linear. O PPG em História também apresentou esse comportamento para 2018, como é analisado a seguir.

4.1.2. PPG – História.

O PPG em História corresponde a 12,1% da Grande Área de Humanidades e chegou à sua fase adulta, como curso de doutorado, em 2018 ao completar 18 anos desde seu início. Diferentemente da grande maioria dos PPGs analisados aqui, o de História apresentou um comportamento de crescimento curvo, ao invés de linear, como é apresentado no Gráfico 2 abaixo.

Gráfico 2 - Fluxo Temporal de Relevância na Pós-Graduação em História.

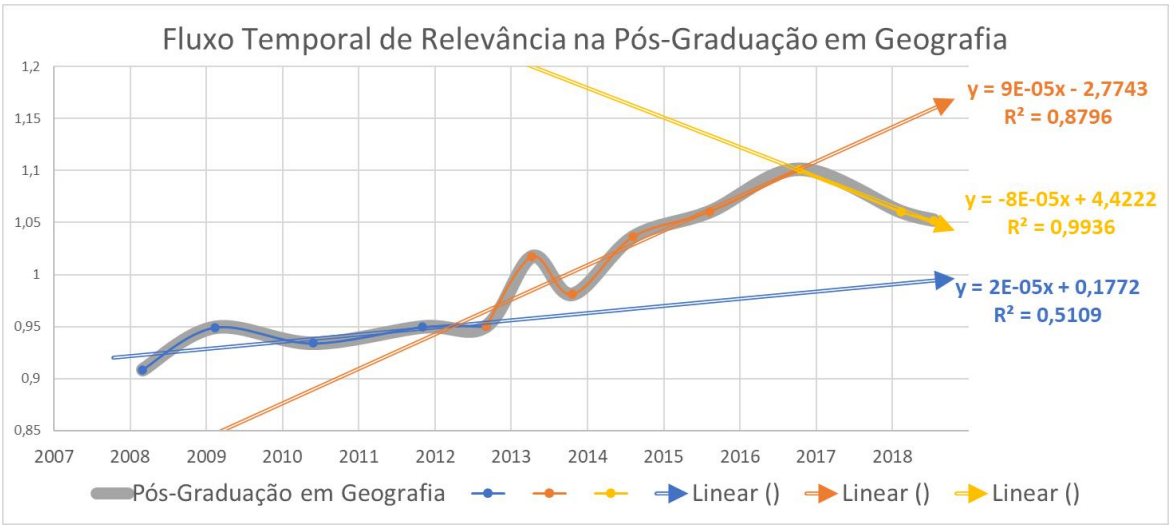


Em virtude desse comportamento curvo, a regressão linear do período mais recente que possui o melhor R^2 foi a partir do final de 2013. O PPG em História apresenta, ao final do período aqui analisado, a sua maior concentração de defesas de teses no ano de 2018, além da média de relevância de tais teses serem as mais altas também. A segmentação temporal necessária para a análise do PPG em História também ocorreu no PPG em Geografia a seguir.

4.1.3. PPG – Geografia.

O PPG em Geografia apresentou uma regressão linear com o R^2 igual a 0,8142. Para encontrar melhores valores, esse PPG foi analisado em três momentos: até 2012, de final de 2012 a 2016 e, por fim, um momento de declínio a partir do final de 2016, conforme Gráfico 3 abaixo.

Gráfico 3 - Fluxo Temporal de Relevância na Pós-Graduação em Geografia.

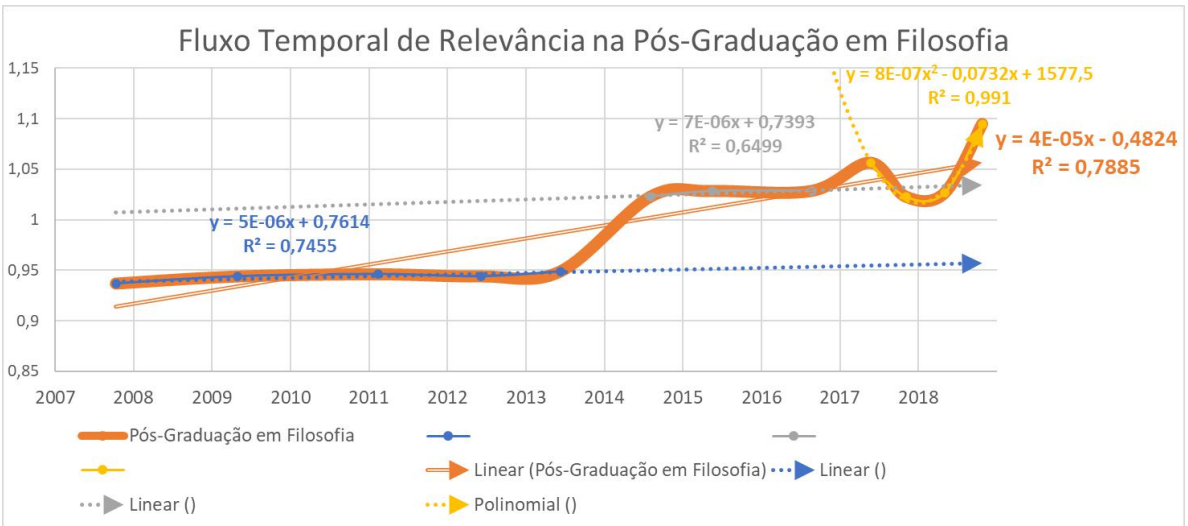


O declínio percebido foi raro dentre as análises de todos PPGs e carece de análises transversais que possam explicá-lo. A nota do curso de doutorado desse PPG é 6 pela CAPES, conforme informação da plataforma Sucupira. Outro PPG que foi analisado com um recorte em três momentos, foi o de Filosofia, conforme subseção a seguir.

4.1.4. PPG – Filosofia.

O curso de doutorado em Filosofia data de 1992 e apresentou uma excepcional estabilidade de valores de relevância nos primeiros anos aqui analisados, até 2013, conforme Gráfico 4 abaixo.

Gráfico 4 - Fluxo Temporal de Relevância na Pós-Graduação em Filosofia.



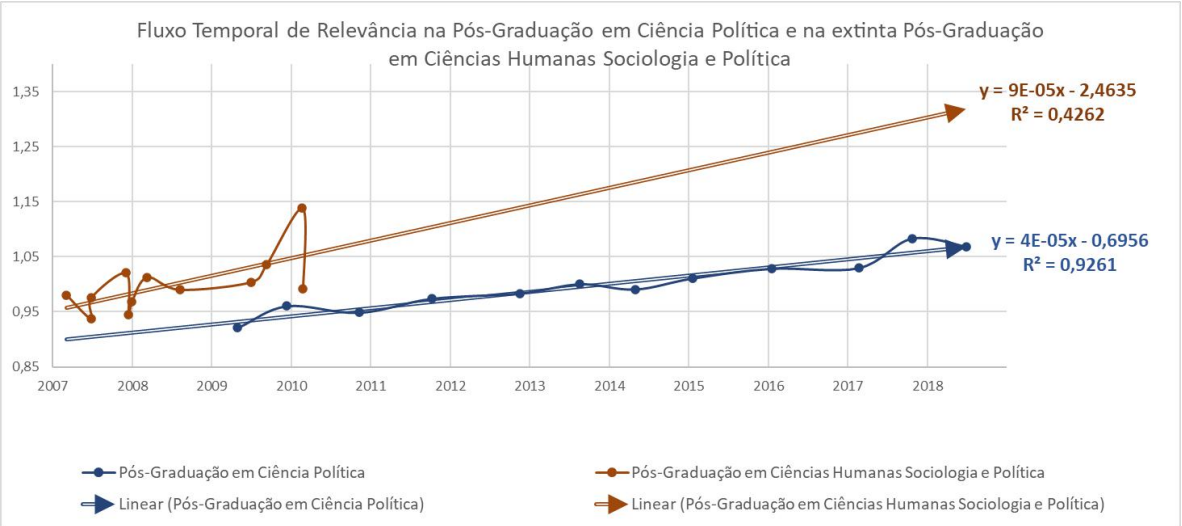
A mesma excepcional estabilidade vista até 2013 ocorre de 2014 a 2016, porém num nível mais elevado. A partir de 2017 os valores de relevância apresentam um comportamento disruptivo, caracterizado por declínio seguido de ascensão, levando-o a atingir seus maiores índices ao final de 2018. Assim como no PPG em Geografia, o conceito CAPES do curso de

doutorado em Filosofia é 6, e carece de informações transversais para uma análise das causas desses comportamentos excepcionais.

4.1.5. PPG – Ciência Política e PPG – Ciências Humanas Sociologia e Política.

O PPG de Ciência Política apresentou um dos melhores comportamentos de estabilidade na variação temporal de relevância de seus termos. No entanto, não foi considerado nas análises comparativas em virtude de o início de seu curso de doutorado ser em 2006, embora, previamente a ele, houve o PPG em Ciências Humanas Sociologia e Política. Optou-se por manter os dados de tais PPGs separados pelo motivo do mais antigo ser mais abrangente e o mais recente, mais específico.

Gráfico 5 - Fluxo Temporal de Relevância na Pós-Graduação em Ciência Política e na extinta Pós-Graduação em Ciências Humanas Sociologia e Política.

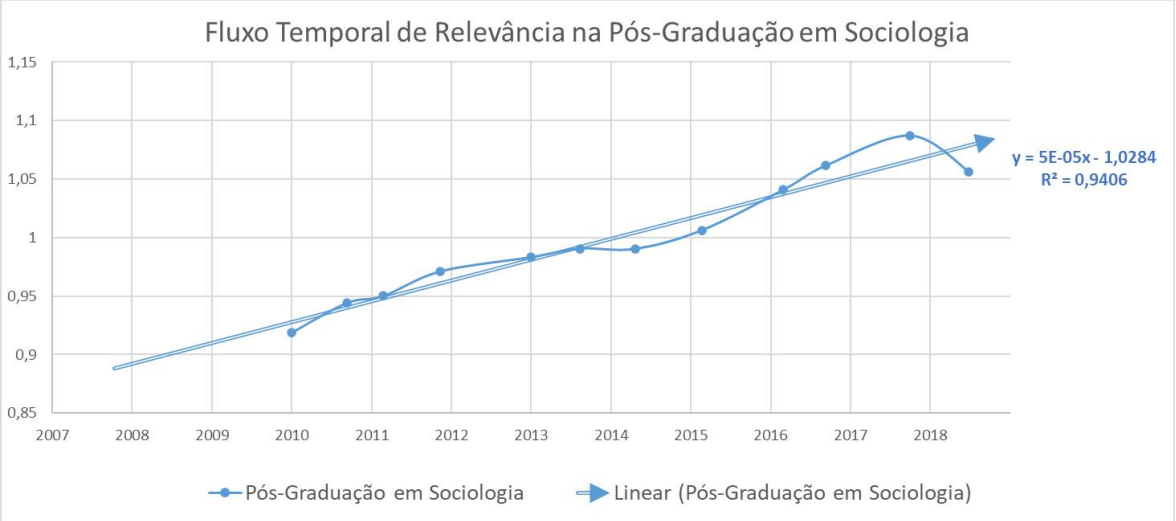


O conceito CAPES para o doutorado no PPG em Ciência Política é atualmente 7, enquanto o anterior, desativado oficialmente em 2017, possuiu a última nota avaliada em 3. Tal desativação está relacionada também à criação do doutorado no PPG em Sociologia, conforme a seguir.

4.1.6. PPG – Sociologia.

O PPG em Sociologia também teve o início de seu doutorado no início de 2007, o que é muito recente para o recorte desta pesquisa, que precisa das teses defendidas a partir de 2007. Logo, assim como para Ciência Política, sua análise é feita isoladamente sem computar nas sínteses por grande área e por Colégio.

Gráfico 6 - Fluxo Temporal de Relevância na Pós-Graduação em Sociologia.



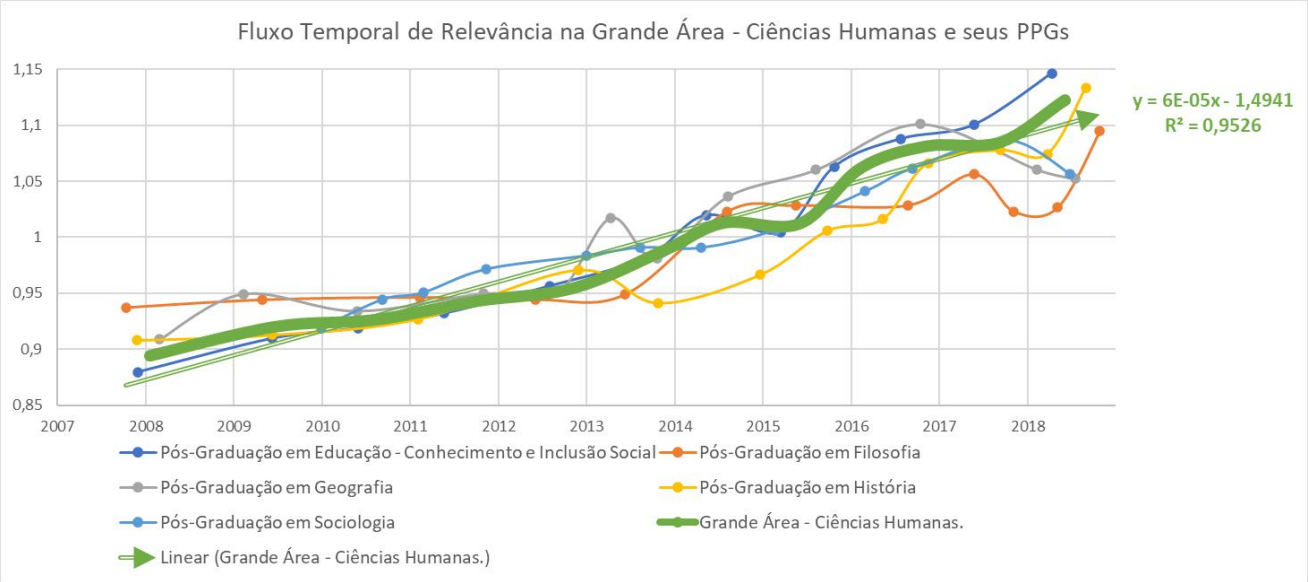
O PPG em Sociologia apresenta um crescimento estável, assim como na Ciência Política. No entanto, é possível perceber um maior crescimento dos valores de relevância a partir de 2015. O leve declínio em 2018 carece de informações adicionais para uma análise mais precisa. Atualmente o curso de doutorado no PPG em Sociologia é avaliado na nota 5 pela CAPES.

4.2. Análise geral de distribuição temporal por PPG

A mais significativa contribuição dessa pesquisa está na análise da variação da amplitude em cada gráfico. É possível perceber uma tendência de crescimento para quase todos os PPGs. Destaca-se que esse fato pode ser considerado como uma possível evolução de cada PPG ao adotar novas terminologias. Somado a isso, há o uso das terminologias pregressas. Cabe para trabalhos futuros, e em cada PPG, uma análise das terminologias específicas e suas ocorrências ao longo do tempo.

O Gráfico 7 apresenta todos os fluxos temporais da grande área de Ciências Humanas e seus PPGs simultaneamente de modo a proporcionar uma visualização de correlações em seus formatos.

Gráfico 7 - Fluxo Temporal de Relevância na Grande Área - Ciências Humanas.



O PPG que mais influenciou na média ponderada dessa grande área foi o de "Educação", pois é o que tem maior quantitativo de teses, 55,2%, conforme Tabela 1. Tal quantitativo influencia numa estabilidade da curva. A qualidade desse resultado geral foi muito além das expectativas, pois dentre todas as regressões, a linear é a mais simples. E somado a isso, um R^2 de 0,9526 é excelente, uma vez que todos os 12 pontos foram incluídos, sem a necessidade de subdividir esse período.

Conforme as Normas Gerais de Pós-Graduação da UFMG, "o Doutorado tem por objetivo desenvolver a capacidade de propor e conduzir, de forma autônoma, pesquisas originais em área específica ou interdisciplinar do conhecimento". Uma vez que parte da pontuação dos termos ocorre pela sua especificidade, podemos inferir que o crescimento ocorre pela adequada condução de "pesquisas originais" nas Ciências Humanas da UFMG em toda a sua trajetória aqui analisada.

5. CONSIDERAÇÕES FINAIS

O critério temporal se mostrou bastante efetivo para ser considerado na pontuação de descritores, que, por sua vez, pode ser aplicada na indexação automática. Podemos concluir que foi possível encontrar uma variação temporal característica da distribuição de valores de termos relevantes ao longo do tempo da produção de textos que contribui como um critério para o processo de sua indexação automática.

A metodologia descrita aqui, com seus 14 passos, apresenta-se atualmente como exclusiva dos autores dessa pesquisa na literatura, sendo fruto parcial de uma pesquisa de doutorado. Igualmente a esse recorte, os autores já possuem resultados para as outras 8 áreas de

conhecimento da UFMG: Ciências da Saúde; Ciências Sociais Aplicadas; Linguística, Letras e Arte; Ciências Exatas e da Terra; Engenharias; Ciências Biológicas; Ciências Agrárias e a denominada como Multidisciplinar.

REFERÊNCIAS

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern Information Retrieval: The concepts and technology behind search**. 2nd ed. Harlow: Pearson Education Limited, 2011.

BORGES, Graciane Bruzinga; LIMA, Gercina Ângela. DESENVOLVIMENTO DE SOFTWARES DE INDEXAÇÃO AUTOMÁTICA: BREVE AVALIAÇÃO DOS PRINCIPAIS CRITÉRIOS. 2015. **XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação** [...]. [*S. l.*: *s. n.*], 2015.

BORKO, Harold. **Indexing concepts and methods**. New York (etc.)London: New York etc.London : Academic Press, 1978, 1978.

BORKO, Harold. Toward a theory of indexing. **Information Processing and Management**, vol. 13, no. 6, p. 355–365, 1977. https://doi.org/10.1016/0306-4573(77)90055-3.

BUSH, Vannevar. As we may think. **The atlantic monthly**, vol. 176, no. 1, p. 101–108, 1945. .

CINTRA, Anna Maria Marques. Elementos de lingüística para estudos de indexação. **Ciência da informação**, vol. 12, no. 1, 1983. .

CORRÊA, Renato Fernandes; DE MIRANDA, Darliane Goes; DE ALMEIDA LIMA, Camila Oliveira; DA SILVA, Tiago José. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ: novas práticas em informação e conhecimento**, vol. 1, no. 1, p. 11–22, 2011. .

DUCHON, Andrew P; MCCORMACK, Robert; SALTER, William J; ALLOPENNA, Paul David; WEIL, Shawn; COLONNA-ROMANO, John; KRAMER, David. **Method and system to predict the likelihood of topics**. [*S. l.*]: Google Patents, 20 Oct. 2015.

KURAMOTO, Hélio. Proposition d'un système de recherche d'information assistée par ordinateur: avec application à la langue portugaise. 1999. .

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, vol. 25, no. 2, p. 182–192, 1996. Available at: http://ridi.ibict.br/handle/123456789/221.

LANCASTER, F W. Indexação e resumos: teoria e prática. Tradução de Antônio Agenor Briquet de Lemos. rev. atual. 2004. .

LUHN, Hans Peter. A statistical approach to mechanized encoding and searching of literary information. **IBM Journal of research and development**, vol. 1, no. 4, p. 309–317, 1957. .

LYONS, John. **Linguagem e linguística: uma introdução**. Rio de Janeiro: LTC - Livros Técnicos e Científicos, 1987.

MAIA, Luiz Claudio Gomes. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. 2008. 158 f. Universidade Federal de Minas Gerais, 2008. Available at: http://hdl.handle.net/1843/ECID-7NXJKZ.

MATHEWS, Litty K; KANMANI, S Deepa. A survey on temporal information retrieval systems. **International Journal of Computer Applications**, vol. 58, no. 4, 2012. .

MESQUITA, Luiz Antônio Lopes; SOUZA, Renato Rocha; PORTO, Renata Maria Abrantes Baracho. Caracterização de testes de oito áreas de conhecimento: uma análise para o desempenho de indexação automática através de sintagmas nominais. 2013. **XIV ENANCIB** [...]. Florianópolis: [*s. n.*], 2013. p. 20.

Available at: http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/2295.

MESQUITA, Luiz Antônio Lopes; SOUZA, Renato Rocha Souza; PORTO, Renata Maria Abrantes Baracho. Noun Phrases in Automatic Indexing: a Structural Analysis of the Distribution of Relevant Terms in Doctoral Theses. 2014. **13th International ISKO Conference - Knowledge Organization in te 21s Century: Between Historical Patterns and Future Prospects.** [...]. Cracow: [*s. n.*], 2014. p. 327–334.

MOOERS, Calvin N. Zatocoding applied to mechanical organization of knowledge. **American documentation**, vol. 2, no. 1, p. 20–32, 1951. https://doi.org/10.1002/ASI.5090020107.

MOULAHI, Bilel; TAMINE, Lynda; YAHIA, Sadok Ben. When time meets information retrieval: Past proposals, current plans and future trends. **Journal of Information Science**, vol. 42, no. 6, p. 725–747, 2016. .

NAVES, MADALENA M LOPES; DIAS, E W. **Análise de assunto: teoria e prática**. [*S. l.*]: Thesaurus Editora, 2007. vol. 3, .

ORTEGA, Cristina Dotta. Relações históricas entre biblioteconomia, documentação e ciência da informação. **DataGramaZero, Rio de Janeiro**, vol. 5, no. 5, p. A03-1001, 2009. .

PERINI, Mário A; FRAIHA, Sigrid; FULGÊNCIO, Lúcia; NETO, Regina Bessa. O SN em português: A hipótese mórfica. **Belo Horizonte: Revista de Estudos de Linguagem-UFMG**, vol. JUlho/Deze, p. 43–56, 1996. .

SARACEVIC, Tefko. Ciência da informação: origem, evolução e relações. **Perspectivas em ciência da informação**, vol. 1, no. 1, p. 41–62, 1996. .

SOUZA, Renato Rocha. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 215 f. UFMG, 2005.

SPARCK JONES, Karen. A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. **Journal of Documentation**, vol. 28, no. 1, p. 11–21, 1972. https://doi.org/10.1108/eb026526.